

УДК 004.62

РАЗРАБОТКА МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ КЛАССИФИКАЦИИ КРЕДИТНЫХ ОРГАНИЗАЦИЙ С ИСПОЛЬЗОВАНИЕМ ДЕРЕВЬЕВ РЕШЕНИЙ И ИХ АНСАМБЛЕЙ

Е.П. Акишина¹, В.В. Иванов^{1,2}, А.В. Крянев^{1,2}, А.С. Приказчикова²*

¹Объединенный институт ядерных исследований, Дубна, 141980, Россия

²Национальный исследовательский ядерный университет «МИФИ», Москва, 115409, Россия

*e-mail: AVKryanev@mephi.ru

Поступила в редакцию: 10.05.2024

После доработки: 19.05.2024

Принята к публикации: 20.05.2024

В статье рассматривается задача применения деревьев решений и их ансамблей (лесов решений) в задаче классификации кредитных организаций, как объектов экономической безопасности. Хотя деревья решений и их ансамбли успешно применяются в банковском секторе, для автоматизированной идентификации неблагонадежных кредитных организаций деревья решений и их ансамбли ранее не использовались. Классификация кредитных организаций проводилась на основе формы банковской отчетности № 101. В результате проведенного анализа были выделены ключевые показатели деятельности кредитных организаций, а именно «Прибыль», «Счета в Банке России», «Ценные бумаги». С учетом этих показателей для модели CART была получена точность классификации, составившая 85 %. Для моделей Random Forest, Adaboost и Xgboost использовались все 23 показателя финансовой отчетности (форма 101), а достигнутая при этом точность составила 83, 80 и 80 %, соответственно, решена актуальная научно-практическая задача – разработаны математические модели, позволяющие идентифицировать высокорисковые кредитные организации и прогнозировать риски отзыва у них лицензий. В ходе их применения идентифицирован список потенциально неблагонадежных кредитных организаций, на которые рекомендуется обратить пристальное внимание органам государственной власти.

Ключевые слова: экономическая безопасность, кредитные организации, автоматизированный мониторинг, дерево решений, лес решений, модель CART, Random Forest, Adaboost, Xgboost.

DOI: 10.26583/vestnik.2024.350

EDN LGWWEH

ВВЕДЕНИЕ

В настоящее время существуют различные методы классификации многомерных объектов, к которым, в частности, можно отнести нейронные сети, метод опорных векторов, метод распознавания образов и др. Однако, как показывает практика, для интерпретации полученных результатов эти методы требуют определенной квалификации и знаний экспертов, а правила извлечения значимой информации (результатов) не всегда очевидны.

С целью решения задачи идентификации высокорисковых кредитных организаций, как объектов экономической безопасности, нами был выбран метод деревьев решений, который позволяет оценить обстановку в финансовом секторе рядовому аналитику, а также визуализиро-

вать отдельные аспекты деятельности кредитных организаций [1, 2].

Деревья решений и их ансамбли успешно применяются в банковском секторе, промышленности, медицине, молекулярной биологии. Вместе с тем для автоматизированной идентификации неблагонадежных кредитных организаций деревья решений и их ансамбли ранее не использовались.

Деревья решений основаны на логических схемах, позволяющих получить решение о классификации объектов с помощью ответов на иерархически организованную систему вопросов. Основная идея данного метода заключается в построении бинарного дерева, в котором каждый узел представляет собой условие на признак, а листья – конечный результат работы алгоритма (в нашем случае принадлежность к определенному классу – благонадежных или

неблагонадежных кредитных организаций). Решающее дерево является бинарным, если каждый узел этого дерева при разбиении имеет только двух потомков. Для того чтобы провести классификацию нового случая, необходимо двигаться по дереву от корня до листа, значение метки класса для которого и будет искомым ответом [1].

ОСОБЕННОСТИ ДЕРЕВА CART

В работе рассматривались три наиболее распространенных вида бинарных деревьев решений: CART (Classification And Regression Tree), C5, CHAID. В табл. 1 представлена информация о метриках качества классификации кредитных организаций с использованием указанных моделей.

Таблица 1. Метрики качества классификации для моделей деревьев

Данные за 2021 г.	C5	CART	CHAID
Общая точность, %	69.03	97.95	93.30
Precision, %	66.87	96.81	90.50
Recall, %	99.70	100.00	99.70
F1, %	80.05	98.38	94.87

По результатам применения указанных моделей (реализованных на языке Python) к исследуемым данным (данные за 2021 г.) было установлено, что максимальная точность классификации кредитных организаций достигается при использовании алгоритма CART, в связи с чем данный алгоритм будет детально рассмотрен далее.

CART является алгоритмом построения бинарного дерева решений – дихотомической классификационной модели [1]. Алгоритм на каждом шаге построения дерева последовательно сравнивает все возможные разбиения для всех показателей и выбирает наилучший из них и наилучшее разбиение для этого показателя.

В качестве меры оценки качества разбиения выборки алгоритм CART использует индекс *Gini*, принимающий значения от 0 до 1. Индекс *Gini* для выборки *A* с *n* классами можно представить следующим образом (1):

$$Gini(A) = 1 - \sum_{i=1}^n p_i^2, \quad (1)$$

где p_i – вероятность класса в выборке *A* [3].

В общем случае задача классификации рассматриваемых объектов формулируется так: для некоторого вектора $\vec{x} = (\{x_1, \dots, x_m\})$, где *m* – число показателей, необходимо провести процедуру принятия решения о принадлежности данного вектора к классу из множества $F = (\{f_1, \dots, f_n\})$, где *n* – число классов, основываясь на обучающей выборке данных.

Для того чтобы реализовать дерево решений CART, необходимо задать его характеристики или гиперпараметры, которые будут определять это дерево. В табл. 2 приведены основные гиперпараметры модели дерева решений CART.

Таблица 2. Основные гиперпараметры модели CART

№	Гиперпараметр	Описание
1	criterion	Функция оценки качества разбиения
2	max_depth	Максимальная глубина дерева
3	min_samples_split	Минимальное количество образцов, необходимое для разделения внутреннего узла
4	min_samples_leaf	Минимальное количество объектов в листьях
5	max_features	Количество показателей, которые следует учитывать при поиске лучшего разделения в узле

По результатам отработки алгоритма CART в Python нами были рассчитаны значения точностей на обучающей и тестовой выборках, составившие 100 и 70 %, соответственно. Так как

точность на тестовой выборке оказалась заметно ниже, чем на обучающей, наша модель дерева решений переобучена. Используем полученную модель в качестве базовой, чтобы ее можно

было сравнивать с другими моделями, к которым мы применим методы борьбы с переобучением, а именно настройку гиперпараметров дерева и использование лесов решений.

НАСТРОЙКА ГИПЕРПАРАМЕТРОВ МОДЕЛИ CART

В качестве первого шага исследуем влияние ключевого гиперпараметра дерева решений CART – максимальной глубины (`max_depth`) на точность классификации (рис. 1). При значе-

нии `max_depth`, равном 4, модель становится оптимальной, т.е. отсутствует переобучение. Она очень хорошо соответствует обучающим данным, а также может быть обобщена на новые данные. После построения дерева решений с `max_depth = 4` точность новой модели не стала существенно выше, чем для базовой модели, однако при этом проблема с переобучением была успешно решена: точность для обучающей выборки составила 78.5 %, для тестовой – 70.4 %.

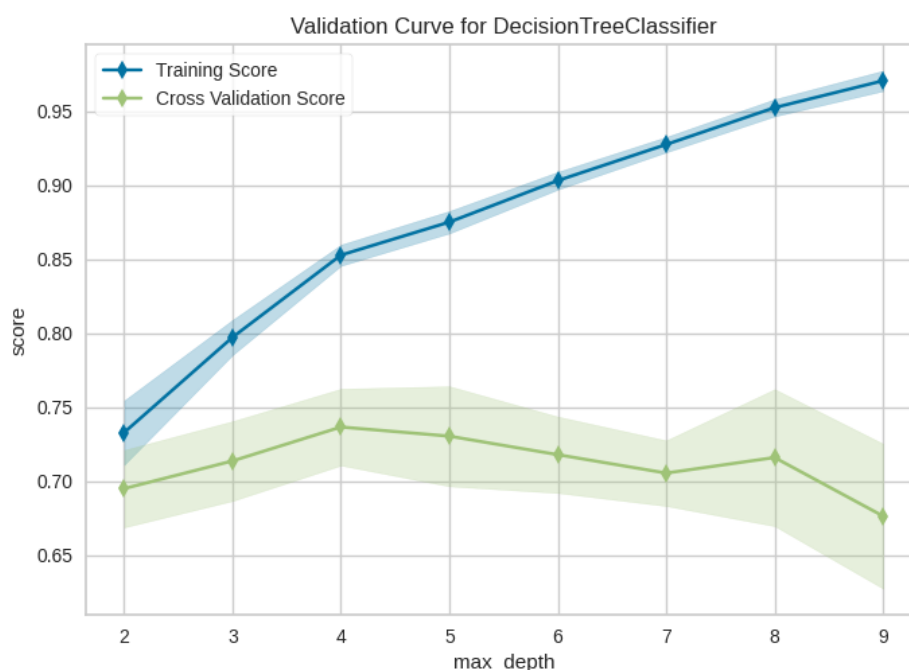


Рис. 1. График зависимости точности модели CART от максимальной глубины дерева

Далее используем поиск комбинации нескольких гиперпараметров для дерева CART, что позволит уменьшить переобучение и оптимизировать точность модели.

В библиотеке `scikit-learn` языка программирования Python реализованы специальные методы подбора оптимальных гиперпараметров, в частности `GridSearchCV` и `RandomisedSearchCV`. Первый метод реализует полный перебор совокупности значений гиперпараметров, а второй – случайный перебор. Использование указанных методов позволяет в полуавтоматическом режиме подбирать наилучшие гиперпараметры, что значительно улучшает качество классификации.

В данной работе мы использовали следующий подход. На первом шаге локализовали область оптимальных значений гиперпараметров с помощью метода `RandomSearchCV`, центром

которой являлась точка в многомерном пространстве гиперпараметров, отвечающая максимальной точности модели CART. На втором шаге мы осуществляли проверку значений гиперпараметров из найденной области с использованием `GridSearchCV`. При таком подходе мы существенно сократили время поиска гиперпараметров и максимально приблизились к их оптимальным значениям.

Результаты поиска оптимальных значений гиперпараметров для дерева CART на тестовых данных приведены в табл. 3.

Как видно из табл. 3, построенное дерево CART обладает точностью классификации на тестовой выборке – 78 %.

С помощью модели CART можно идентифицировать показатели, которые в первую очередь влияют на отнесение объекта к одному из классов: благонадежные кредитные организации

РАЗРАБОТКА МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ КЛАССИФИКАЦИИ КРЕДИТНЫХ ОРГАНИЗАЦИЙ С ИСПОЛЬЗОВАНИЕМ ДЕРЕВЬЕВ РЕШЕНИЙ И ИХ АНСАМБЛЕЙ

(класс «0») и неблагонадежные (класс «1»). В связи с чем в Python была реализована оценка значимости показателей. Наиболее важными для модели CART оказались «Прибыль», «Счета в Банке России» и «Ценные бумаги». На основе выделенных показателей была построена новая модель дерева решений CART, точность

которой увеличилась на 7 % по сравнению с исходной и составила 85 %. На рис. 2 представлена кривая операционных характеристик для модели дерева CART и рассчитано значение AUC – площадь под кривой.

Таблица 3. Оптимальные гиперпараметры для дерева CART

Гиперпараметр CART	Поисковый диапазон	Оптимальное значение
Максимальная глубина	(3, 11)	6
Максимальное количество признаков	(4, 24)	11
Максимальное количество листьев	(3, 30)	11
Минимальное количество объектов в листе	(1, 21)	10
Минимальное количество объектов в узле для его разделения	(2, 21)	2
Минимальное уменьшение неоднородности	[0.0001, 0.0005, 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.5, 0.7]	0.001
Точность модели	78%	

Перейдем к анализу структуры построенного дерева CART. Рассмотрим листы (результатирующие классы) дерева CART, индекс *Gini* которых не более 0.2. В листовых элементах 1–4 дерева CART содержатся правильно классифици-

рованных 96, 91, 100 и 94 % неблагонадежных объектов, соответственно.

На рис. 3, в качестве примера, представлена характеристика первого листового элемента.

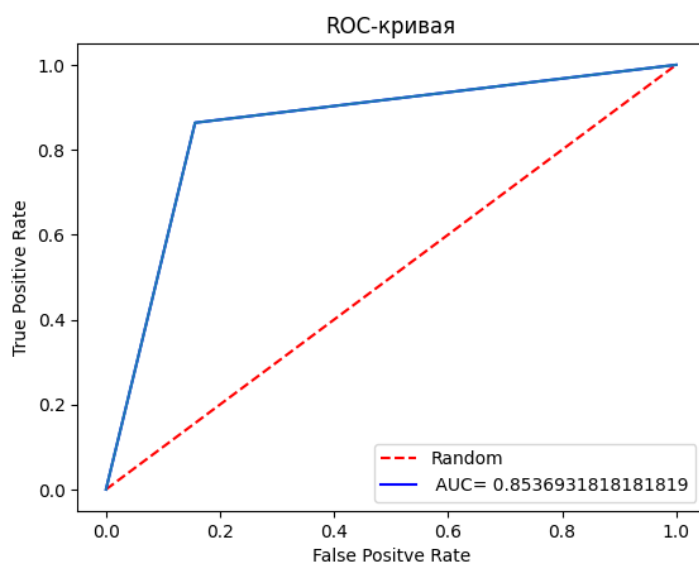
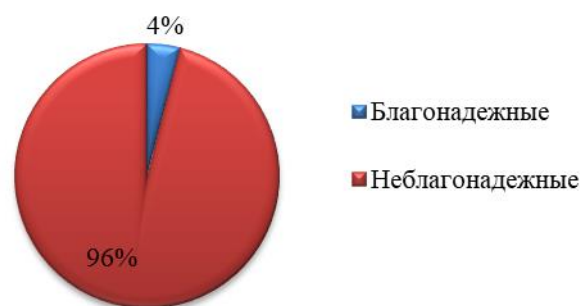


Рис. 2. Кривая операционных характеристик для модели дерева CART



Путь до листа № 1:
 Прибыль ≤ -7282000
 Ценные бумаги ≤ 775474496
 Счета в ЦБ ≤ 316499488
 Прибыль ≤ -27401500

Рис. 3. Характеристика первого листового элемента дерева CART

ЛЕСА РЕШЕНИЙ: МЕТОД RANDOM FOREST

Одним из самых перспективных методов, позволяющим понизить уровень переобучения, является ансамблирование деревьев решений в леса решений. Леса решений строят множество деревьев решений и усредняют их результаты. Рассмотрим леса, показавшие наибольшую эффективность при решении задач финансового мониторинга: Random Forest, Adaboost и Xgboost. Все леса рассчитывались по исходным 23 показателям.

Random Forest или случайный лес – один из способов объединения деревьев решений в ансамбль. В задаче классификации Random Forest выбирает решение по большинству результатов, выданных совокупностью деревьев, сочетая в себе идеи метода бэггинга [4] и метода случайных подпространств [5]. В методе Random

Forest, как и в методе бэггинга, обучение происходит независимо друг от друга на разных подмножествах обучающей выборки, что решает проблему построения одинаковых деревьев на одном и том же множестве данных. Как и в методе случайных подпространств, в методе Random Forest деревья обучаются по случайным подмножествам показателей.

При реализации случайного леса важнейшим этапом является определение его гиперпараметров. При этом в качестве критерия разделения узла при решении задачи классификации может быть использован критерий *Gini*. В табл. 4 представлены результаты поиска оптимальных гиперпараметров для модели Random Forest.

На рис. 4 представлена кривая операционных характеристик для Random Forest и рассчитана площадь под кривой – AUC.

Таблица 4. Оптимальные гиперпараметры Random Forest

Гиперпараметр Random Forest	Поисковой диапазон	Оптимальное значение
Количество деревьев	(100, 1000)	300
Максимальная глубина	(3, 10)	6
Максимальное количество признаков	(4, 23)	17
Максимальное количество листьев	(3, 15)	5
Минимальное количество объектов в листе	(1, 10)	1
Минимальное количество объектов в узле для его разделения	(2, 10)	2
Точность модели	83 %	

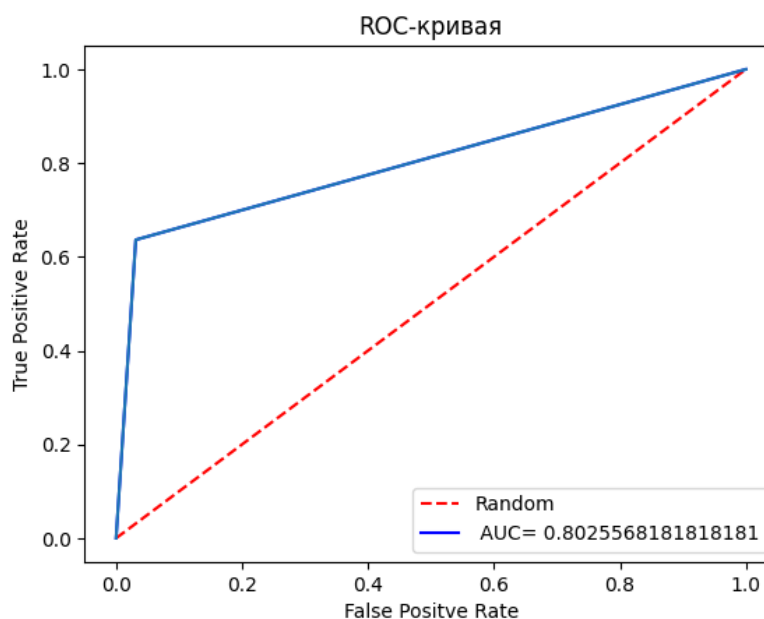


Рис. 4. Кривая операционных характеристик для Random Forest

ЛЕСА РЕШЕНИЙ: МЕТОДЫ ADABOOST И XGBOOST

В основу методов Adaboost и Xgboost заложена идея бустинга, заключающаяся в последовательном обучении каждой следующей модели с использованием данных об ошибках предыдущих моделей.

Adaboost является первой моделью бустинга. Adaboost минимизирует экспоненциальную функцию потерь, что может сделать алгоритм чувствительным к выбросам. Adaboost выполняет взвешивание анализируемых объектов, которые были неправильно классифицированы ранее, и на следующих этапах производит «работу над ошибками» по выявленным инцидентам [6].

Метод Xgboost включает в себя идеи метода градиентного спуска и бустинга. На каждом этапе вычисляется градиент функции потерь по предсказаниям предыдущих классификаторов. Следующий классификатор корректирует предсказания предыдущих. С алгоритмом Xgboost можно использовать любую дифференцируемую функцию потерь, а следовательно, Xgboost более устойчив к выбросам, чем Adaboost [7, 8].

В табл. 5 и 6 представлены результаты поиска оптимальных значений гиперпараметров для моделей Adaboost и Xgboost, соответственно. На рис. 5 и 6 представлены кривые операционных характеристик и рассчитана площадь под кривой – AUC для моделей Adaboost и Xgboost соответственно.

Таблица 5. Оптимальные гиперпараметры Adaboost

Гиперпараметр Adaboost	Поисковой диапазон	Оптимальное значение
Количество деревьев	[100, 200, 300, 500, 1000, 1500, 2000, 3000, 5000]	2000
Алгоритм	['SAMME', 'SAMME.R']	SAMME.R
Максимальная глубина	(3, 11)	4
Максимальное количество листьев	(4, 25)	18
Максимальное количество признаков	(4, 24)	8
Минимальное количество объектов в листе	(1, 11)	1
Минимальное количество объектов в узле для его разделения	(2, 11)	6
Скорость обучения	[0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1.0]	0.1
Точность модели	80 %	

Таблица 6. Оптимальные гиперпараметры для Xgboost

Гиперпараметр Xgboost	Поисковой диапазон	Оптимальное значение
Количество деревьев	[100, 200, 300, 500, 1000, 1500, 2000, 3000, 5000]	3000
Метод построения дерева	['exact', 'approx', 'hist']	hist
Максимальная глубина	(3, 11)	6
Максимальное количество листьев	(4, 25)	10
Скорость обучения	[0.0001, 0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3]	0.1
Параметр gamma	[0.0, 0.1, 0.5, 1, 1.5, 2]	1.5
Параметр reg_lambda	(0, 0.8)	0.6
Параметр reg_alpha	(0, 0.6)	0.5
Доля выборки для обучения	(0.5, 1)	0.8
Доля признаков для обучения	(0.5, 1)	0.5
Точность модели	80 %	

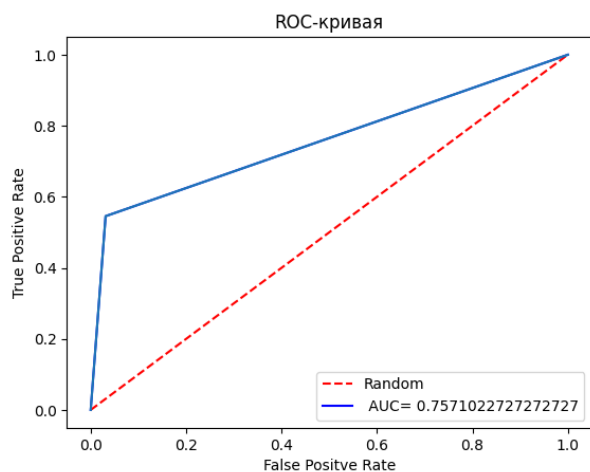


Рис. 5. Кривая операционных характеристик для Adaboost

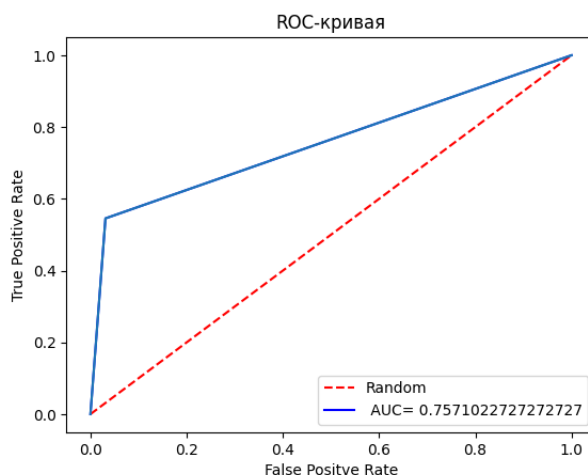


Рис. 6. Кривая операционных характеристик для Xgboost

На рис. 7 представлена гистограмма, которая демонстрирует эффект от проведения поиска оптимальных значений гиперпараметров модели дерева CART и моделей лесов решений Random Forest, Adaboost, Xgboost. Видно, что максимальный результат достигнут при оптимизации дерева CART при условии построения модели на трех параметрах (15 %). Следующей по величине прироста точности классификации идет модель Random Forest (13 %), далее Adaboost (6 %) и, наконец, Xgboost (4 %).

В табл. 7 представлены итоговые метрики качества модели дерева CART и моделей лесов

решений Random Forest, Adaboost, Xgboost. Общая точность (Accuracy) для модели CART, реализованной на трех показателях, оказалась максимальной и составила 85 %. Высокие значения метрик Recall, F1-score и AUC для модели CART также свидетельствуют о целесообразности ее применения. Однако максимальное значение точности классификации неблагонадежных объектов (Precision) было достигнуто при использовании модели Random Forest и составило 93 %.

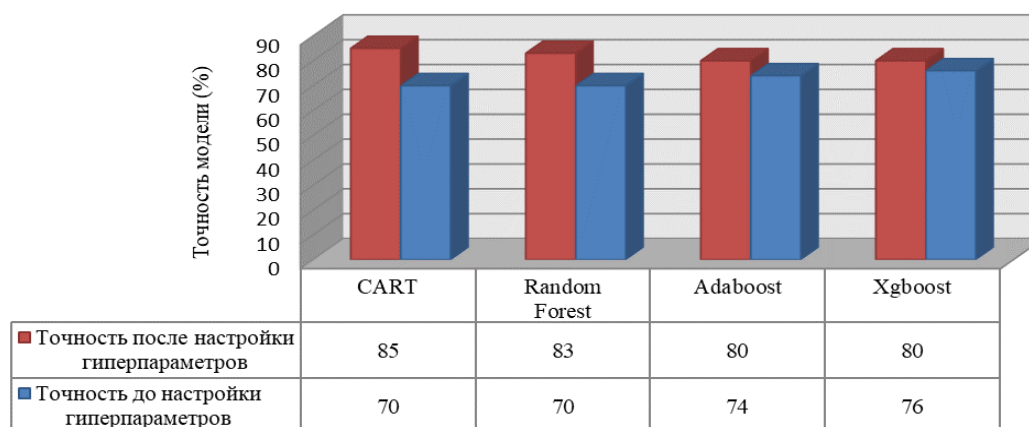


Рис. 7. Сравнительный анализ точностей моделей до и после поиска оптимальных гиперпараметров

Таблица 7. Итоговые метрики качества моделей деревьев решений и их ансамблей

Метрика	CART		Random Forest		Adaboost		Xgboost	
	1	0	1	0	1	0	1	0
Accuracy	85	85	83	83	80	80	80	80
Precision	79	90	93	79	92	76	92	76
Recall	86	84	64	97	55	98	56	97
F1-score	83	87	76	87	69	86	69	85
AUC	85	85	80	80	76	76	76	76

АНАЛИЗ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ
И ВЫВОДЫ

В настоящей работе исследована проблема идентификации высокорисковых кредитных организаций с помощью моделей дерева решений CART и лесов решений Random Forest, Adaboost и Xgboost. В результате проведенного анализа были выделены ключевые показатели деятельности кредитных организаций, а именно «Прибыль», «Счета в Банке России», «Ценные бумаги». С учетом этих показателей для модели CART была получена максимальная точность классификации, составившая 85 %. Для моделей Random Forest, Adaboost и Xgboost использовались все 23 показателя финансовой отчетности (форма 101), а достигнутая при этом точность составила 83, 80 и 80 %, соответственно. Таким образом, нами была решена актуальная научно-практическая задача – разработаны математические модели, позволяющие идентифицировать высокорисковые кредитные организации и прогнозировать риски отзыва у них лицензий.

СПИСОК ЛИТЕРАТУРЫ

1. Han J., Kamber M., Pei J. Data mining. Concepts and techniques. 3th ed. Elsevier Inc., 2012. 740 p.
2. Акишина Е.П., Иванов В.В., А. Крянев А.В., Приказчикова А.С. Многомерный анализ данных в задачах прогнозирования попадания кредитных организаций в зону риска // Вестник НИЯУ МИФИ, 2024. Т. 13, № 1. С. 22–29. DOI: 10.26583/vestnik.2024.302
3. Гмурман, В.Е. Руководство к решению задач по теории вероятностей и математической статистике. М.: Высш. Школа, 2003. 400 с.
4. Breiman L. Random Forests // Machine Learning, 2001. V. 45. No. 1. P. 5–32.
5. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning : Data Mining, Inference, and Prediction. 2th ed. Springer – Verlag, 2009. 746 p. ISBN: 978-0-387-84857-0.
6. Mishra A. Fraud Detection: A Study of AdaBoost Classifier and K-Means Clustering // SSRN. 2021. № 2 (16). P. 1–9. DOI: 10.21203/rs.3.rs-247874/v1.
7. Vassallo D. Vella V., Ellul J. Application of Gradient Boosting Algorithms for Anti-money Laundering in Crypto currencies // SN Computer Science, 2021. № 2 (3). P. 142-157. DOI: 10.1007/s42979-021-00558-z.
8. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Computer Science Machine Learning, 2016 (v1). URL: <https://arxiv.org/abs/1603.02754> (дата обращения: 08.02.2024).

Vestnik Natsional'nogo Issledovatel'skogo Yadernogo Universiteta «MIFI», 2024, vol. 13, no. 4, pp. 242–250

**MATHEMATICAL MODELS DEVELOPMENT FOR CREDIT INSTITUTIONS
CLASSIFICATION USING DECISION TREES AND THEIR ENSEMBLES**

E.P. Akishina¹, V.V. Ivanov^{1,2}, A.V. Kryanev^{1,2*}, A.S. Prikazchikova²

¹Joint Institute for Nuclear Research, Dubna, 141980 Russia,

²National Research Nuclear University «MEPhI», Moscow, 115409 Russia,

*e-mail: AVKryanev@mephi.ru

Received May 10, 2024; revised May 19, 2024; accepted May 20, 2024

The article examines the problem of using decision trees and their ensembles (decision forests) in the problem of classifying credit institutions as objects of economic security. Although decision trees and their ensembles have been successfully used in the banking sector, decision trees and their ensembles have not previously been used for automated identification of unreliable credit institutions. The classification of credit institutions was carried out on the basis of bank reporting form No. 101. As a result of the analysis, key performance indicators of credit institutions were identified, namely «Profit», «Accounts with the Bank of Russia», «Securities». Taking into account these indicators, the classification accuracy of 85 % was obtained for the CART model. For the Random Forest, Adaboost, and Xgboost models, all 23 financial statement measures (Form 101) were used, and the accuracy achieved was 83, 80, and 80 %, respectively. an urgent scientific and practical problem has been solved - mathematical models have been developed that make it possible to identify high-risk credit organizations and predict the risks of revo-

cation of their licenses. During their application, a list of potentially unreliable credit institutions was identified, to which it is recommended that government authorities pay close attention.

Keywords: economic security, credit organizations, automated monitoring, decision tree, decision forest, CART model, Random Forest, Adaboost, Xgboost.

REFERENCES

1. *Han J., Kamber M., Pei J.* Data mining. Concepts and techniques 3th ed. Elsevier Inc., 2012. 740 p.
2. *Akishina E.P., Ivanov V.V., Kryanev A.V., Pri-kazchikova A.S.* Mnogomernyj analiz dannyh v zadachah prognozirovaniya popadaniya kreditnyh organizacij v zonu riska [Multidimensional data analysis in problems of predicting whether credit institutions will fall into the risk zone]. Vestnik NIYAU MIFI, 2024. Vol. 13. No. 1. Pp. 22–29. DOI: 10.26583/vestnik.2024.302
3. *Gmurman, V.E.* Rukovodstvo k resheniyu zadach po teorii veroyatnostej i matematicheskoy statistike. [Guide to solving problems in probability theory and mathematical statistics]. Moscow, Vyssh. Shkola Publ., 2003. 400 p.
4. *Breiman L.* Random Forests Machine Learning, 2001. Vol. 45. No. 1. Pp. 5–32.
5. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2th ed. Springer-Verlag, 2009. 746 p. ISBN: 978-0-387-84857-0.
6. *Mishra A.* Fraud Detection: A Study of AdaBoost Classifier and K-Means Clustering. SSRN. 2021. No. 2 (16). Pp. 1–9. DOI: 10.21203/rs.3.rs-247874/v1.
7. *Vassallo D. Vella V., Ellul J.* Application of Gradient Boosting Algorithms for Anti-money Laundering in Crypto currencies. SN Computer Science, 2021. No. 2 (3). Pp. 142–157. DOI: 10.1007/s42979-021-00558-z.
8. *Chen T., Guestrin C.* XGBoost: A Scalable Tree Boosting System. Computer Science Machine Learning, 2016 (v1). Available at: <https://arxiv.org/abs/1603.02754> (accessed: 08.02.2024).